16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

SPEECH ENHANCEMENT BASED ON A HYBRID *A PRIORI* SIGNAL-TO-NOISE RATIO (SNR) ESTIMATOR AND A SELF-ADAPTIVE LAGRANGE MULTIPLIER

Md. Jahangir Alam[1], Sid-Ahmed Selouani[2] and Douglas O'Shaughnessy[3]

[1,3]INRS-Energie-Matériaux-Télécommunications, Université du Québec, Montréal QC H5A 1K6, Canada
[2]Université de Moncton, Campus de Shippagan NB E8S 1P6, Canada
Phone : +1-514-875-1266, Fax : +1-514-875-0344, email : [1]alam@emt.inrs.ca , [3]dougo@emt.inrs.ca, selouani@umcs.ca[2]
Web: www.emt.inrs.ca

ABSTRACT

Speech enhancement techniques, using spectral subtraction, have the drawback of generating an annoying residual noise with musical character. An accurate estimate of the a priori SNR is critical for eliminating musical noise. In this paper, for an accurate estimate of the a priori SNR we have proposed a hybrid a priori SNR estimator and a self-adaptive Lagrange multiplier with Wiener denoising technique. Objective evaluations showed that the proposed method performed better than the Decision-Directed (DD) approach.

Keyword: *a priori* SNR, hybrid, decision directed, Wiener filter, speech enhancement

1. INTRODUCTION

Most voice communication systems are designed for processing noise-free speech. Speech signals used as an input to these systems are often degraded by additive noise. Speech enhancement has therefore attracted a great deal of research interest to reduce the noise level in noisy speech.

Although most speech enhancement techniques improve speech quality, they often suffer from an annoying artifact called musical noise, caused by randomly spaced spectral peaks that come and go in each frame, and at random frequencies. The randomly spaced peaks are due to inaccurate and large variance estimates of the spectra of the noise and noisy signals [3].

Many approaches on noisy speech enhancement have been investigated in the last few decades. Among them, the minimum mean square error (MMSE) estimation approach is one of the most popular speech enhancement techniques, as it can reduce the musical noise that is a common feature existing in other approaches [1]. The dominant point behind the reduction of musical noise by the MMSE approach is the DD approach for the *a priori* SNR estimation, but the *a priori* SNR follows the *a posteriori* SNR with a frame delay [5]. Since the spectral gain function depends on the estimated *a priori* SNR, spectral gain computed at the current frame matches the previous frame and therefore the performance of the speech enhancement technique is degraded. We have proposed a new method called a hybrid *a priori* SNR estimation approach, which solves this problem while maintaining the advantages of the DD approach. The proposed method shows improved performance over the DD method when applied with or without a self-adaptive Lagrange multiplier.

The organization of this paper is as follows: Section 2 of this paper presents a description of the well known Wiener denoising method. A description of the proposed hybrid *a priori* SNR estimation approach is made in section 3. Section 4 provides a description of the self-adaptive Lagrange multiplier. Performance evaluations and the conclusion of this paper are made in section 5 and section 6, respectively.

2. WIENER DENOISING METHOD

Let the distorted signal be expressed as

$$y(n) = x(n) + d(n)\,, \tag{1}$$

where $x(n)$ is the clean signal and $d(n)$ is the additive random noise signal, uncorrelated with the original signal. If at the mth frame and kth frequency bin $Y(m,k)$, $X(m,k)$ and $D(m,k)$ represent the spectral components of $y(n)$, $x(n)$ and $d(n)$, respectively, then the distorted signal in the transformed domain is

$$Y(m,k) = X(m,k) + D(m,k)\,. \tag{2}$$

An estimate $\widehat{X}(m,k)$ of $X(m,k)$ is given by

$$\widehat{X}(m,k) = H(m,k)Y(m,k)\,, \tag{3}$$

where $H(m,k)$ is the noise suppression gain (denoising filter), which is a function of *a priori* SNR and *a posteriori* SNR, given by

$$H(m,k) = \left(\frac{\xi(m,k)}{\mu + \xi(m,k)}\right)^{\beta}, \tag{4}$$

where μ is a constant, β is the order of the filter and $\xi(m,k)$ is the *a priori* SNR. If μ=1 and β=1/2 then (4) corresponds to power spectrum filtering. For a generalized Wiener Filter $\beta = 1$.

The first parameter of the noise suppression rule is the *a posteriori* SNR given by

$$\gamma(m,k) = \frac{|Y(m,k)|^2}{\Gamma_d(m,k)}\,, \tag{5}$$

where $\Gamma_d(m,k) = E\{|D(m,k)|^2\}$ is the noise power spectrum estimated during speech pauses using the classical recursive relation:

$$\Gamma_d(m,k) = \lambda_D \Gamma_d(m-1,k) + (1-\lambda_D)|Y(m,k)|^2, \quad (6)$$

where $0 \le \lambda_D \le 1$ is the smoothing factor. In this paper we have chosen $\lambda_D = 0.9$ for all cases. $E\{.\}$ is the expectation operator.

The *a priori* SNR, which is the second parameter of the noise suppression rule, is expressed as

$$\xi(m,k) = \frac{\Gamma_x(m,k)}{\Gamma_d(m,k)}, \quad (7)$$

where $\Gamma_x(m,k) = E\{|X(m,k)|^2\}$.

The *instantaneous* SNR [6] can be defined as

$$\vartheta(m,k) = \frac{|Y(m,k)|^2}{\Gamma_d(m,k)} - 1. \quad (8)$$

The temporal-domain denoised speech is obtained with the following relation

$$\hat{x}(n) = IFFT(|\hat{X}(m,k)|.e^{j\arg(Y(m,k))}). \quad (9)$$

3. ESTIMATION OF *A PRIORI* SNR

An important parameter of numerous speech enhancement techniques is the *a priori* SNR. Although most speech enhancement techniques improve speech quality, they suffer from an annoying artifact called musical noise caused by randomly spaced spectral peaks that come and go in each frame, and at random frequencies. The randomly spaced peaks are due to the inaccurate estimate of the *a priori* SNR [3]. An accurate estimate of the a priori SNR is critical for eliminating musical noise.

3.1 DECISION-DIRECTED APPROACH

A widely used method to determine the *a priori* SNR from distorted speech is the decision-directed (DD) approach. In [2] the DD approach was defined as a linear combination of (7) and (8). With a weighting parameter α that is constrained to be $0 < \alpha < 1$, the linear combination results in

$$\xi(m,k) = E\{\alpha \frac{|X(m,k)|^2}{\Gamma_d(m,k)} + (1-\alpha)\vartheta(m,k)\}. \quad (10)$$

However, as this expression is hard to implement in practice, approximations were made. This led to the following expression:

$$\hat{\xi}_{DD}(m,k) = \alpha \frac{|H_{DD}(m-1,k)Y(m-1,k)|^2}{\Gamma_d(m,k)} + (1-\alpha)P'[\vartheta(m,k)], \quad (11)$$

where $P'[x] = x$ if $x \ge 0$ and $P'[x] = 0$ otherwise. In this paper we have chosen $\alpha = 0.98$ by the simulations and informal listening tests. The multiplicative gain function for this approach is

$$H_{DD}(m,k) = \frac{\hat{\xi}_{DD}(m,k)}{\mu(m,k) + \hat{\xi}_{DD}(m,k)}. \quad (12)$$

Then the enhanced speech spectrum is obtained using (3). $\mu(m,k)$ is described in section 4.

An important characteristic of the DD approach is the dependency on previously enhanced frames, which results in biased estimates of the *a priori* SNR during speech transitions. This method results in a significant elimination of musical noise.

3.2 PROPOSED HYBRID *A PRIORI* SNR ESTIMATOR

In the conventional DD approach the weighting factor α is of constant value (close to unity), the speech spectrum estimated in the previous frame is used to estimate the current *a priori* SNR and the *a priori* SNR follows the *a posteriori* SNR with a delay of one frame when the *a posteriori* SNR exhibits an abrupt increase [5]. This frame delay produces undesired gain distortion and thus generates audible distortion during abrupt transient periods. The musical noise is significantly reduced during noise frames when the weighting factor α increases but during the speech onset periods the speech signal could be distorted.

To suppress the problem of the decision-directed approach while maintaining its benefits, we propose a hybrid *a priori* SNR estimation method, which can provide fast response to an abrupt increase in the speech signal without introducing musical noise. We have used the DD approach when the change in *a posteriori* SNR between the current frame and the previous frame is greater than a certain threshold; otherwise the modified approach is applied so that the estimated *a priori* SNR can appropriately follow the shape of the original speech during transient periods. The algorithm is discussed below:

If $\Delta\gamma(k) > Thrld$

$$\alpha_h(m,k) = \alpha,$$

$$\hat{\xi}_h(m,k) = \alpha \frac{|H_h(m-1,k)Y(m-1,k)|^2}{\Gamma_d(m,k)} + (1-\alpha)P'[\vartheta(m,k)],$$

else

$$\alpha_h(m,k) = \alpha_m(m,k),$$

$$\hat{\xi}_h(m,k) = \alpha_m(m,k) \frac{|H_h(m-1,k)Y(m-1,k)|^2}{\Gamma_d(m,k)} + (1-\alpha_m(m,k))P'[\vartheta(m,k)],$$

where $0 < \alpha_m(m,k) < 1$, is the modified weighting factor, based on the previous *a posteriori* SNR and is given by the following relation:

$$\alpha_m(m,k) = \frac{1}{1 + \left(\frac{\Delta\gamma(k)}{\max(\gamma(m,k), \gamma(m-1,k)) + 1} \right)^2}, \quad (13)$$

where $\Delta\gamma(k) = |\gamma(m,k) - \gamma(m-1,k)|$, the threshold is $Thrld = E\{\gamma(m,k)\}$, $k = 1,2,3,..,K$ is the spectral bin index and $m = 1,2,3,..,M$ is the frame index, K is the frame length and M is the number of frames. The estimate of the *a priori* SNR in the proposed approach is given by:

$$\hat{\xi}_h(m,k) = \alpha_h(m,k) \frac{|H_h(m-1,k)Y(m-1,k)|^2}{\Gamma_d(m,k)} + (1-\alpha_h(m,k))P'[\vartheta(m,k)], \quad (14)$$

where $H_h(m,k)$ is the spectral gain for the proposed approach and is given by

$$H_h(m,k) = \frac{\hat{\xi}_h(m,k)}{\hat{\xi}_h(m,k) + \mu(m,k)}. \qquad (15)$$

Figure 1 depicts the block diagram of the proposed hybrid *a priori* SNR estimator for the *m*th frame with frame length *K*. Figure 2 represents the variation of the average weighting factor (per frame) of both approaches. Figure 3 represents the variation of the *a priori* SNR of the proposed approach and that of the DD approach with the *a posteriori* SNR. The proposed hybrid approach efficiently avoids the delay generated by the DD approach and the estimated *a priori* SNR resembles the *a posteriori* SNR during speech onset periods.

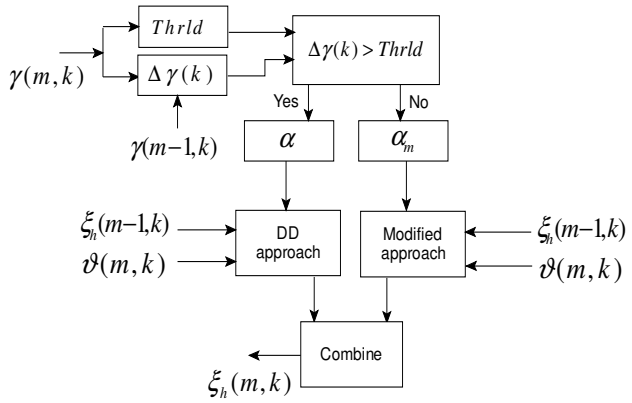

Figure 1 Block diagram of the proposed hybrid *a priori* SNR estimation approach for the *m*th frame with frame length *K*.

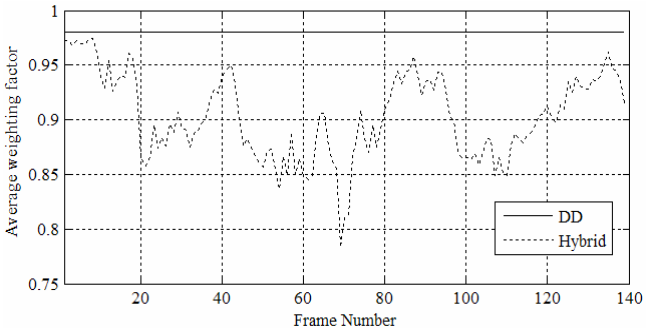

Figure 2 Average (frequency-averaged) value of the weighting factor (α and $\alpha_h(m,k)$) of the both approaches. Subway Noise, SNR=10 dB.

4. SELF-ADAPTIVE LAGRANGE MULTIPLIER

The role of the Lagrange multiplier [3] is the same as that of the over-subtraction factor in [4]. The Lagrange multiplier $\mu(m,k)$ ($\mu(m,k) \geq 1$) in (4), (12) and (16) controls the trade off between speech distortion and residual noise. A large μ would produce more speech distortion with less residual noise. Conversely a small μ would produce a smaller amount of speech distortion with more residual noise. Since the speech signal will mask the noise in the speech-dominated frames, we would like to reduce the speech distortion in those frames and would like to reduce residual noise in noise-dominated frames.

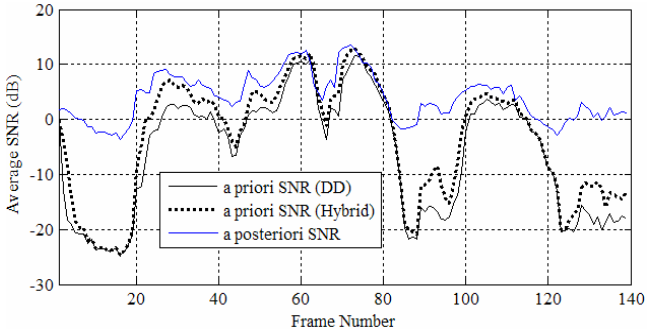

Figure 3 Variations of the *a priori* SNR of the DD approach and that of the proposed approach with the *a posteriori* SNR. Subway Noise, SNR= 5 dB.

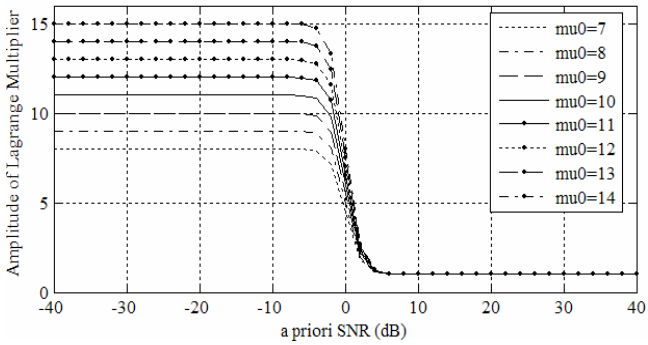

Figure 4 Variation of μ with the *a priori* SNR (dB) at different values of μ_0.

We have made the value of μ dependent on the estimated *a priori* SNR, $\hat{\xi}_h(m,k)$ and it is expressed by the following relation

$$\mu(m,k) = 1 + \mu_0 \left(1 - \frac{1}{1 + e^{-\xi_{dB}(m,k)}} \right), \qquad (16)$$

where $\xi_{dB}(m,k) = 10\log_{10}(\hat{\xi}_h(m,k))$, and μ_0 is a constant chosen experimentally. We have used $\mu_0 = 9$ on the basis of simulations. Figure 4 shows the variation of the self-adaptive Lagrange multiplier, $\mu(m,k)$ with the *a priori* SNR (dB) at different values of μ_0.

5. PERFORMANCE EVALUATION AND DISCUSSION

A. OBJECTIVE QUALITY MEASURES

To measure the quality of the enhanced signal we have used the Log Likelihood Ratio (LLR), the Log Spectral Distance (LSD), and the Segmental SNR [7]. All three measures show high correlation with informal listening tests.

The most popular class of the time domain measures is the segmental SNR. It is well known that segmental SNR is more accurate in indicating the speech distortion than the overall SNR. The frame based segmental SNR is formed by averaging frame level SNR estimates and is defined by

$$SegSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\sum_{i=mK}^{i=mK+K-1}\frac{x^2(i)}{(x(i)-\hat{x}(i))^2}, \qquad (17)$$

where $x(i)$ is the original speech, $\hat{x}(i)$ is the processed speech reproduced by a speech processing system, K is the length of the segment and M is the number of segments in the speech signal. The higher value of the segmental SNR indicates the weaker speech distortions.

The LLR is referred to as the Itakura distance measure. It is defined as

$$LLR = \log\left(\frac{\bar{a}_x R_{\hat{x}} \bar{a}_x^T}{\bar{a}_{\hat{x}} R_{\hat{x}} \bar{a}_{\hat{x}}^T}\right), \qquad (18)$$

where $\bar{a}_x$ is the LPC coefficient vector $\{1, -a_x(1), -a_x(2), ..., -a_x(p)\}$ for the original speech signal $x(n)$, $\bar{a}_{\hat{x}}$ is the LPC coefficient vector $\{1, -a_{\hat{x}}(1), -a_{\hat{x}}(2), ..., -a_{\hat{x}}(p)\}$ for the processed speech $\hat{x}(n)$, p is the order of LPC coefficient, and $R_{\hat{x}}$ is the autocorrelation matrix for the processed speech. The lower the LLR measure for an enhanced speech, the better is its perceived quality.

The Log Spectral Distance is defined as

$$LSD = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\sum_{k=mK}^{k=mK+K-1}\left[\hat{X}(k) - X(k)\right]^2, \qquad (19)$$

where $X(k)$ is the power spectra of the original speech, $\hat{X}(k)$ is the power spectra of the processed speech reproduced by a speech processing system, K is the length of the segment and M is the number of segments in the speech signal. The higher LSD reflects the stronger speech distortions.

B. EXPERIMENTAL RESULTS

In this section, the performance of the proposed approach is tested for speech enhancement and compared to that of the DD approach with and without the proposed self-adaptive Lagrange multiplier. In order to evaluate the performance of the proposed hybrid *a priori* SNR estimation approach described in section 3.2, we conducted extensive objective quality tests under various noisy environments. The frame sizes were chosen to be 256 samples (32 msec) long with 40% overlap; a sampling frequency of 8 kHz and a hamming window were applied. To evaluate and compare the performance of the *a priori* SNR estimators, we carried out simulations with the *TEST A* and the *TEST B* databases of Aurora [8]. Speech signals were degraded with five types of noise at global SNR levels of 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. The noises were N1 (Subway noise), N2 (Babble Noise) from the *TEST A* database, N1 (Restaurant Noise), N2 (Street Noise), and N3 (Airport Noise) from the *TEST B* database.

Table 1, Table 2, and Table 3 represent the Average segmental SNR, the LSD, and the LLR, respectively, of the enhanced signals at different input SNR levels. Experimental results show that the proposed hybrid *a priori* SNR estimator (with and without $\mu(m,k)$)-based method performed better than the conventional DD approach-based method. Figure 5 represents the spectrograms of the clean signal and enhanced signals obtained with the DD approach and the proposed hybrid approach. The speech spectrograms provide more accurate information about the residual noise and speech distortion than the corresponding time domain waveforms. We compared the spectrograms for each of the methods and confirmed a reduction of the residual noise and speech distortion. Speech spectrograms presented in the figure use a hamming window of length 256 samples with 50% overlap and the noisy signals include Street noise with SNR= 5 dB. Experimental results, plotted spectrograms, and informal listening tests show that the proposed technique performs better in all tested objective quality measures; it does not introduce additional speech distortion, and results in significant reduction of the musical noise phenomenon.

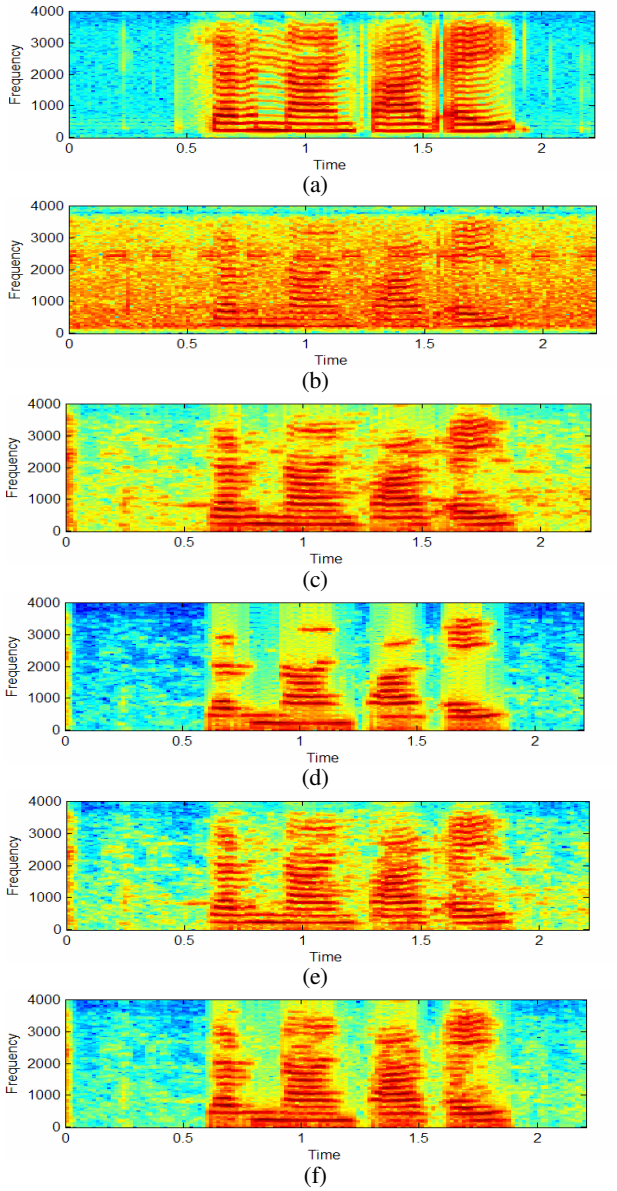

Figure 5 Speech Spectrograms, Street Noise, SNR=5 dB:(a) clean signal, (b) noisy signal, and enhanced signals obtained using (c) the DD approach, (d) the DD approach with $\mu(m,k)$, (e) the hybrid approach, and (f) the hybrid approach with $\mu(m,k)$.

Table 1 Average segmental SNR

Noise Type	Input SNR (dB)	DD with $\mu(m,k)$	DD	Hybrid with $\mu(m,k)$	Hybrid
Subway	0	-0.949	-1.2091	**0.159**	-0.979
	5	1.375	2.210	**2.515**	2.211
	10	4.626	5.109	**5.9205**	5.528
	15	7.372	8.038	**8.740**	8.346
	20	10.371	10.636	**10.901**	10.547
Babble	0	-2.938	-2.430	**-1.393**	-2.014
	5	0.312	0.057	**0.938**	0.320
	10	3.134	2.527	**4.607**	3.329
	15	5.409	5.247	**6.368**	5.639
	20	8.263	7.612	**9.222**	8.385
Restaurant	0	-3.216	-2.300	**-0.545**	-1.108
	5	1.958	2.406	**3.993**	3.451
	10	4.239	4.878	**5.207**	4.863
	15	7.352	7.940	7.991	**8.006**
	20	10.188	10.345	**10.490**	10.359
Street	0	-3.200	-2.751	-2.770	**-2.621**
	5	-0.483	-0.101	**1.386**	0.451
	10	3.778	2.812	**4.886**	3.370
	15	5.451	4.905	**6.993**	5.677
	20	7.457	7.435	**9.300**	7.811
Airport	0	-1.330	-1.058	**-0.582**	-1.078
	5	1.617	1.450	**2.620**	1.497
	10	4.289	4.179	**4.980**	4.627
	15	8.187	8.222	**8.900**	8.536
	20	10.744	10.835	**11.668**	11.281

Table 2 Log Spectral Distance (LSD)

Noise Type	Input SNR (dB)	DD with $\mu(m,k)$	DD	Hybrid with $\mu(m,k)$	Hybrid
Subway	0	2.014	2.127	**1.987**	2.112
	5	1.994	1.703	1.738	**1.622**
	10	1.775	1.705	**1.483**	1.592
	15	1.666	1.403	1.369	**1.291**
	20	1.434	1.188	1.314	**1.186**
Babble	0	2.039	2.088	**1.958**	2.020
	5	1.736	1.745	**1.559**	1.649
	10	1.639	1.540	1.577	**1.452**
	15	1.574	1.492	1.441	**1.420**
	20	1.742	1.381	1.778	**1.392**
Restaurant	0	2.341	2.105	2.524	**2.141**
	5	1.838	1.606	1.835	**1.516**
	10	1.810	1.532	1.703	**1.421**
	15	1.538	1.350	1.394	**1.252**
	20	1.435	1.224	1.327	**1.187**
Street	0	2.195	2.278	**1.908**	2.136
	5	1.856	1.846	**1.663**	1.696
	10	1.805	1.577	1.698	**1.503**
	15	1.602	1.428	1.443	**1.332**
	20	1.588	1.249	1.368	**1.191**
Airport	0	1.873	1.764	1.844	**1.740**
	5	1.794	1.542	1.804	**1.460**
	10	1.652	1.453	1.558	**1.437**
	15	1.664	1.371	1.507	**1.269**
	20	1.688	1.340	1.602	**1.267**

Table 3 Log Likelihood Ratio (LLR)

Noise Type	Input SNR (dB)	DD with $\mu(m,k)$	DD	Hybrid with $\mu(m,k)$	Hybrid
Subway	0	1.540	1.532	**1.289**	1.321
	5	1.341	0.986	0.913	**0.786**
	10	1.197	0.873	0.616	**0.652**
	15	1.060	0.613	0.525	**0.482**
	20	0.550	0.450	0.428	**0.398**
Babble	0	1.466	1.311	**1.218**	1.145
	5	1.178	1.063	0.981	**0.943**
	10	0.974	0.900	0.776	**0.771**
	15	0.989	0.990	**0.814**	0.855
	20	0.973	0.943	**0.864**	0.871
Restaurant	0	1.765	1.364	1.652	**1.294**
	5	0.963	0.822	0.809	**0.710**
	10	1.062	0.984	0.826	**0.766**
	15	0.835	0.846	**0.628**	0.687
	20	0.7211	0.736	**0.559**	0.589
Street	0	1.518	1.309	1.163	**1.011**
	5	1.309	1.056	0.935	**0.815**
	10	1.096	0.969	**0.854**	0.864
	15	0.749	0.733	**0.633**	0.635
	20	0.714	0.618	**0.516**	0.529
Airport	0	1.517	1.325	1.375	**1.240**
	5	1.234	1.202	0.998	**0.925**
	10	1.382	0.868	0.906	**0.802**
	15	0.825	0.688	0.592	**0.541**
	20	1.083	1.018	0.829	**0.826**

6. CONCLUSION

In this paper we have proposed a hybrid *a priori* SNR estimator which avoids the delay problem of the DD approach while keeping its advantages and a self-adaptive Lagrange multiplier for the wiener denoising technique. Performance evaluations of the proposed approach are carried out using three objective quality measures [7]. Simulation results show that the proposed algorithm possesses better performance for speech enhancement in various noisy environments than that of the conventional DD approach.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", Proc. IEEE, vol. 67, pp. 1586–1604, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 32, pp. 1109–1121, 1984.

[3] Yi Hu, Phillips C. Loizu, "Speech Enhancement based on wavelet thresholding the Multitaper Spectrum," IEEE Trans. on Speech and Audio Processing, vol. 12, no. 1,pp. 59-67, January 2004.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1979, pp. 208-211.

[5] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech and Audio Processing, vol. 2, no. 1, pp. 345–349, April 1994.

[6] Cyril Plapous, Claude Marro, Laurent Mauuary and Pascal Scalart, "A Two Step Noise Reduction Technique", IEEE Trans. Acoustic, Speech and Signal Processing, vol. 1, pp. 289–292, 2004.

[7] S. R. Quackenbush, T. P. Barnwell and M. A. Clements, *Objective measures of speech quality*. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[8] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy environments", ISCA ITRW ASR, September 2000.